APPENDIX B

<u>Pseudo Code Algorithm - Identify Variables As Categorical Or Continuous</u>
```
        If (fieldtype = Boolean) then vartype = categorical
        If (fieldtype = float) then vartype = continuous
        If (fieldtype = text and C > Xmax) then variable is dropped
        If (fieldtype = text and C ≤ Xmax) then vartype = categorical
        If ((fieldtype = integer or long integer) and C ≤ Cmax) then vartype = categorical
        If ((fieldtype = integer or long integer) and C > Cmax) then
                If (Pearson's r > Rmin) then
                        // Correlation between the target and this predictor
                        vartype = continuous
                Else
                        For each category c
                                If (Nc < Nmin) then
                                        Recode record as missing
                                        //Note that this actually creates a new variable
                        End For
                        Recalculate C
                        If (C = 0) then
                                vartype = continuous
                                Quit
                        Else If (0 < C ≤ Cmax) then
                                vartype = categorical
                                Quit
                        Else (C > Cmax)
                                Sort bins in ascending order on those unique values
                                Do until (MAX(p-value) < Tmin or C <= Cmax)
                                        For each adjacent pair of bins A and B
                                                Construct the associated target subsets TA and TB
                                                Perform T-test on TA and TB and calculate the
                                                corresponding p-value
                                        End For
                                        Find MAX(p-value)
                                        // Note that MAX(p-value) = the maximum p-value across
                                        all //adjacent pairs of bins
                                        If (MAX(p-value) >= Tmin) then
                                                Combine corresponding bins A and B.
                                                C = C-1
                                End Do
                                Recalculate C
                                If C ≤ Cmax then
                                        vartype = categorical
                                Else
                                        vartype = continuous
```

51

// Note that in this case we use the original variable both to
//build and deploy the model – undo possible
collapses.

End All

where:

C = the count of the number of unique values ('bins') within a variable, exclusive of missing values;

N_c = the count of the number of records in the Cth bin;

Records = the count of the number of records;

Target = A continuous variable;

Xmax = the upper bound on the number of categories permitted for a text-valued categorical variable. The default value is 25;

Cmax = the upper bound on the number of categories permitted for an integer-valued categorical variable. The default value is 10;

Nmin = the minimum number of observations within a category. The default value is 5;

Rmin = the minimum level of Pearson's r for a continuous variable to be considered a "strong predictor." The default value is 0.5;

Tmin = the cutoff significance level from the T-test to collapse adjacent cells. The default value is 0.05.

It is understood that the default values given above are exemplary only and may be adjusted in order to modify the criteria for identifying categorical variables.

Methods of performing T-test and p-value calculations are well known in the art. Given two data sets A and B, the standard error of the difference of the means can be estimated by the following formula:

$$S_D = \sqrt{\frac{\sum_A (x_i - \bar{x}_A)^2 + \sum_B (x_i - \bar{x}_B)^2}{size(A) + size(B) - 2} + \left(\frac{1}{size(A)} + \frac{1}{size(B)}\right)}$$

where t is computed by

$$t = \frac{\bar{x}_A - \bar{x}_B}{S_D}$$

Finally, the significance of the t (p-value) for a distribution with size(A) + size(B)-2 degree freedom is evaluated by the incomplete beta function